A Comprehensive Exploitation of Instance Selection Methods for Automatic Text Classification

Washington Cunha
DCC-UFMG
Belo Horizonte, Brazil
washingtoncunha@dcc.ufmg.br

Leonardo Rocha
DCOMP-UFSJ
São João del-Rei, Brazil
lcrocha@ufsj.edu.br

Marcos André Gonçalves
DCC-UFMG
Belo Horizonte, Brazil
mgoncalv@dcc.ufmg.br

***Abstract*—Progress in Natural Language Processing (NLP) has been dictated by the rule of more: more data, more computing power and more complexity, best exemplified by the Large Language Models. However, training (or fine-tuning) large dense models for specific applications usually requires significant amounts of computing resources. This Ph.D. dissertation focuses on an under-investigated NLP data engineering technique, whose potential is enormous in the current scenario known as Instance Selection (IS). The IS goal is to reduce the training set size by removing noisy or redundant instances while maintaining the effectiveness of the trained models and reducing the training process cost. We provide a comprehensive and scientifically sound comparison of IS methods applied to an essential NLP task – Automatic Text Classification (ATC), considering several classification solutions and many datasets. Our findings reveal a significant untapped potential for IS solutions. We also propose two novel IS solutions that are noise-oriented and redundancy-aware, specifically designed for large datasets and transformer architectures. Our final solution achieved an average reduction of 41% in training sets, while maintaining the same levels of effectiveness in all datasets. Importantly, our solutions demonstrated speedup improvements of 1.67x (up to 2.46x), making them scalable for datasets with hundreds of thousands of documents. This thesis falls under the Neural Systems and Machine Learning topic of the CBIC-CTD call for papers.**

***Index Terms*—Instance Selection, Text Classification, Deep Learning, Sustainable and Responsible AI.**

Dissertation available at: http://hdl.handle.net/1843/76441
Publications available at: http://bit.ly/3WvX1T5

I. Introduction

The rapid data growth on the Web, social network platforms, companies, and governmental institutions has made organizing and retrieving content extremely challenging. Automatic Text Classification (ATC) offers a solution to this problem by mapping textual documents into predefined semantic categories. Accurate ATC models have become crucial for many emerging applications [1], such as spam, fake news and hate speech detection, relevance feedback, sentiment and product review analysis, to cite just a few. As a supervised task, ATC benefits from applications generating large volumes of *labeled data*, such as social networks (e.g., X, Facebook, WhatsApp). Crowdsourcing and soft labeling [2] further reduce the costs of acquiring labeled data. Thus, labeling has become less of an issue, while the abundance of labeled data is.

According to Andrew Ng [3], the success of Transformer-based architectures, the state-of-the-art (SOTA) in ATC, best exemplified by Small and Large Language Models (SLMs and LLMs) such as RoBERTa and Llama 4, is due to extensive pre-training on massive datasets (e.g., 45PB for GPT-4) and the adaptability of pre-trained models via fine-tuning. This approach enables faster task-specific training compared to starting from scratch [4]. However, fine-tuning remains resource-intensive. Despite being faster than full training, it still requires significant computational power. For instance, fine-tuning the SLM XLNet in the MEDLINE dataset took 80 GPU hours in our experiments. Resource limitations in companies and research groups also restrict experimentation with such models. For instance, in our PhD dissertation alone, we ran over 5,000 experiments that took approximately 5,600 hours in a specialized (GPU-based) architecture. Reducing financial, computational, and environmental costs is crucial, given the significant energy consumption and carbon emissions associated with generating and using (large) language models.

II. Motivation and Objective

Given increasing data volumes, re-training demands, and environmental concerns, proposing scalable and cost-effective NLP and ATC strategies has become essential. These include creating efficient deep learning algorithms, using advanced hardware, or improving data preprocessing techniques. The recent success and real-world impact, including financial, of DeepSeek [5], which matched or surpassed the effectiveness of SOTA LLMs while reducing computational demands, highlights the importance of the trade-off effectiveness vs. cost to the research and practitioner communities. This PhD dissertation focuses exactly on this trade-off, one of the SBC 2025-2035 Grand Challenges on Computer Science on AI Sustainability, from a *data engineering perspective*, by reducing training computational costs and carbon footprint without compromising performance. In particular, we focus on Instance Selection (IS), an understudied (in NLP and ATC at least), yet promising, set of techniques and growing research area [6]–[8], focused on selecting the most representative instances (documents) for a training set [9]. The intuition behind IS is to remove potentially noisy or redundant instances

This work was supported by CNPq, CAPES, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR), FAPEMIG, AWS, Google, NVIDIA, Finep, CIIASaúde, and FAPESP.

from the original training set and improve performance in terms of total training time while keeping or even improving effectiveness. IS methods have three main concomitant goals: *(i) to reduce the number of instances by selecting the most representative ones; (ii) to maintain (or even improve) effectiveness by removing noise and redundancy;* and *(iii) to reduce the total time for applying an end-to-end model (from preprocessing to model training to deployment (test).*

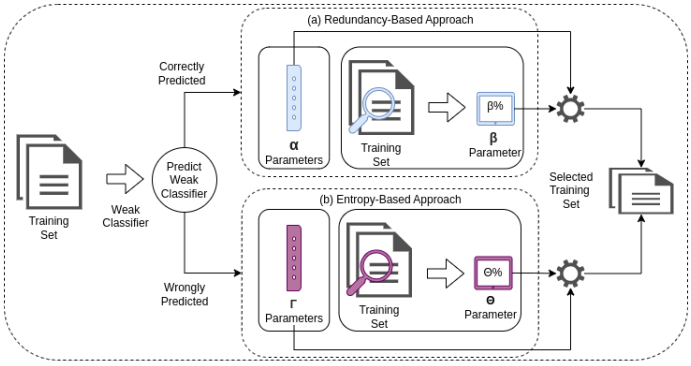

Fig. 1. Bi-objective Instance Selection Framework

III. Summary of the Proposed Solution

The main contributions of our PhD dissertation are fourfold: (i) a comprehensive survey of the IS methods applied to ATC, including (ii) an extended IS taxonomy; and (iii-iv) two novel state-of-the-art (SOTA) IS approaches applied to NLP/ATC. Due to space limitations, we focus on the latter two, noticing that our article on the ACM Computing Surveys [1], derived from the dissertation, has been highly cited (52 citations as of April/2025). First, we proposed **E2SC** [10] (Figure 1 – in blue), a two-step IS framework aimed at large datasets with a special focus on transformer-based architectures. E2SC's first step assigns a probability to each instance being removed from the training set (α parameter). We adopted an exact KNN model solution to estimate the probability of removing (training) instances, as KNN is considered a calibrated [11] and computationally cheap (fast) classifier. Our first hypothesis (H1) was that *high confidence (if the model is calibrated to the correct class known in training) positively correlates with redundancy for the sake of building a stronger classification model*. In the second E2SC step, we estimate a near-optimal reduction rate that does not degrade the Transformer's effectiveness (β parameter) by employing a validation set and a weak but fast classifier. Our second hypothesis (H2) was that *we can estimate the effectiveness of a robust model through the analysis and variation of selection rates in a weaker model*. Again, we explored KNN to gather evidence for this hypothesis by introducing an iterative method that statistically compared, using a validation set, the KNN model's effectiveness without any data reduction against the model with iterative data reduction rates. In this way, we could estimate a reduction rate that did not affect the KNN model's effectiveness.

E2SC focused only on *redundancy*. Despite excellent results regarding the trade-off effectiveness-efficiency-reduction, other aspects such as **noise** – defined as instances incorrectly labeled by humans [12] as well as outliers that do not contribute to model learning – were not explored in our first solution. To fill this gap, we proposed **biO-IS** [13], built on top of E2SC, aimed at simultaneously removing redundant and noisy instances. **biO-IS** has three main components: (i) a weak classifier; (ii) a redundancy-based approach; and (iii) an entropy-based approach (the lower part of Figure 1 – in purple). We departed from E2SC, considering the Logistic Regression (LR) as the calibrated weak classifier instead of KNN – in further experiments described in the dissertation, LR proved to be the best classifier for effectiveness and calibration. To address the second objective of noise removal, we proposed a new step to be combined with our previous IS solution based on entropy (Γ parameter), as well as a novel iterative process to estimate near-optimum reduction rates (θ parameter). Considering wrongly predicted instances by the weak classifier, the main objective is to assign a probability to each of them being removed from the training set based on the probability of the instance being noisy. For this, we proposed using entropy as a proxy to determine the reduction behavior for the sake of training a stronger model. Accordingly, the proposed biO-IS framework provides a comprehensive solution to address redundancy and noise removal simultaneously.

IV. Evaluation

We compared our proposals with 13 of the most robust SOTA IS baseline methods in the ATC domain based on our systematic literature review, considering 22 datasets and 7 SOTA small and large language models (including BERT, RoBERTa, Llama, among others). Our experimental evaluation showed that **E2SC** was able to reduce the size of the training sets by 29% on average while maintaining the same levels of effectiveness in almost all datasets, with speedups of 1.37x on average. The framework scaled to large datasets, reducing them by up to 40% while statistically maintaining the same effectiveness with speedups of 1.70x. E2SC focused only on redundancy, however. **biO-IS**, in turn, extended E2SC, being capable of removing, besides redundancy, also noisy instances in up to 66.6%. biO-IS managed to significantly reduce the training sets (by 40.1% on average; up to 60%) while maintaining the same effectiveness levels in **all** of the considered datasets. biO-IS was also capable of consistently producing speedups of 1.67x on average (maximum of 2.46x). No baseline, not even E2SC, was able to achieve results with this level of quality, considering all tripod criteria. biO-IS improved over E2SC by 41% regarding reduction rate and from 1.42x to 1.67x (on average) regarding speedup, being the current state-of-the-art (SOTA) in Instance Selection applied to NLP.

V. Contributions and State-of-the-art advancement

In the Ph.D. dissertation, we conducted a rigorous comparative study of classical and SOTA IS methods applied to ATC. The study evaluated tradeoffs among reduction, effectiveness, and cost, motivated by the rising costs of new ATC solutions due to contextual embeddings, Transformer architectures, and increasing data volumes. Our findings show, contrary to common beliefs, Transformers often require representative - not large - data to perform well in ATC. Overall, IS techniques effectively reduced training set sizes without compromising effectiveness. The previous SOTA in IS fell short of meeting all tripod criteria simultaneously, underscoring the need for more efficient, scalable IS solutions, especially in big data scenarios. To address these challenges and fill the gaps found in the literature, we proposed two novel IS methods focused on redundancy (only) and noise (in conjunction with redundancy). Extensive experimental evaluation confirmed our hypotheses: *small and large language models can be trained with less data without sacrificing effectiveness*. This not only enables cost savings but also contributes to reducing carbon emissions. Such experimental evaluation established our solutions as the current SOTA IS applied to NLP. Such promising results instill hope for a more sustainable (green) and efficient NLP future, where advancements in IS techniques can produce environmental and economic benefits. Indeed, the proposed IS solutions pave the way for developing environmentally sustainable and computationally efficient NLP systems, essential for Brazil's leadership in responsible AI.

VI. Scientific Production

This PhD directly resulted in 4 journal papers (including ACM TOIS, IP&M, and ACM Computing Surveys), 2 conference papers (SIGIR and ICTIR), and open-source software releases. More specifically, our work on IS has been published in the main Machine Learning and Computational Intelligence venues:

1) **Cunha, Washington**, et al. "A Noise-Oriented and Redundancy-Aware Instance Selection Framework." **ACM Transactions on Information Systems** (ACM TOIS) 43.2 (2025): 1-33 – h5-index: 48.0
2) **Cunha, Washington**, et al. "A quantum annealing instance selection approach for efficient and effective transformer fine-tuning." International Conference on Theory of Information Retrieval **ICTIR'24**. p. 205-214 – h5-index: 24.0
3) **Cunha, Washington**, et al. "An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification." **ACM SIGIR** 2023. p. 665-674 – h5-index: 103.0.
4) **Cunha, Washington**, et al. "A Comparative Survey of Instance Selection Methods applied to Non-Neural and Transformer-Based Text Classification." **ACM Computing Surveys** 55.13s (2023): 1-52 – h5-index: 157.0
5) **Cunha, Washington**, et al. "On the cost-effectiveness of neural and non-neural approaches and representations for text classification:A comprehensive comparative study."**IP&M** 58.3 (2021): 102481 - h5-index: 114.0
6) **Cunha, Washington**, et al. "Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling." **IP&M** 57.4 (2020): 102263 – h5-index: 114.0

In addition, the work on my Ph.D. gave me the opportunity and the expertise to contribute as a co-author to several other published journal articles (8 in total), listed below:

1) Bittencourt, G., **Cunha, W** et al. (2025). [14] On representation learning-based methods for effective, Review-Aware Recommender Systems (RARSs): Recent Advances, Experimental Comparative Analysis, Discussions, and New Directions. **ACM Computing Surveys**. - h5-index: 157; Impact Factor: 23.8
2) França, C., **Cunha, W** et al. (2024). [15] On representation learning-based methods for effective, efficient, and scalable code retrieval. **Neurocomputing**. - h5-index: 136; Imp. Fac.: 5.5
3) Andrade, C., **Cunha, W** (2023). [16] On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". **IP&M**. h5-index: 96; IF: 7.4
4) Zanotto, B. S., **Cunha, W** et al. (2021). [17] Pcv50 automatic classification of electronic health records for a value-based program through machine learning. **Value in Health**. – h5-index: 57.0; IF: 4.9
5) Zanotto, B. S., **Cunha, W** et al. (2021). [18] Stroke outcome measurements from electronic medical records: Cross-sectional study on the effectiveness of classifiers. **JMIR Med** – h5-idx: 52.0; IF:4.9
6) Viegas, F., **Cunha, W** et al. (2024). [19] Exploiting contextual embeddings in hierarchical topic modeling and investigating the limits of the current evaluation metrics.**Comp. Linguistics**- h5-index:38; IF:3.7
7) Felix, L. et al., **Cunha, W** (2024). [20] Why are you traveling? Inferring trip profiles from online reviews and domain-knowledge. **Online Social Networks and Media**. – h5-index: 28.0; IF: 4.4
8) Viegas, F., **Cunha, W** et al. (2024). [21] Pipelining semantic expansion and noise filtering for sentiment analysis of short documents – clusent method. **Journal on Interactive Systems**. – h5-index: 9.0

Ideas, insights, and methods of our dissertation also contributed to other 20 conference papers, including: **GDCOMP** [22], **ACL'25** [23] (h5-index: 215), **ACL'20** [24] (h5-index: 215), **SIGIR'25** [25] (h5-index: 103), **CIKM** [26] (h5-index: 91), **WSDM** [27] (h5-index: 77), **ECIR'25** [28] (h5-index: 42), **CoNLL** [29] (h5-index: 39), **WebSci'25** [30] (h5-index: 38), **ENIAC'24** [31] (h5-index: 16), **WebMedia** [23], [32]–[35] (h5-index: 13), **IIR** [36] (h5-index: 7.0), **SBRC'25** [37] (h5-index: 7.0), **SBBD** [38] (h5-index: 7.0).

The combined h5-index of all above publications is **1501**. The respective papers have received so far more than **646** (according to Google Scholar[1]). For more details, we refer the reader to http://bit.ly/3WvX1T5. The h-index of the PhD is 12, which is high for someone who just obtained his PhD title.

Technical Production: We make the documented code of our proposed methods as well as all compared methods (IS and classifiers), datasets (and fold splits), available to the community for replication and comparisons on GitHub [2]. We believe making all these artifacts is very useful for reproducibility and comparison of future methods.

VII. Awards and Achievements

During the Ph.D., the student has received several awards:

1) Finalist of the Thesis and Dissertation Contest of the Brazilian Computer Society (**CTD-SBC'25**)
2) Best reviewer on **SIGIR**'24 and **ACL**'24 conferences;
3) **CTIC'24** and **CTIC'19**: co-advisor of undergraduate work ranked among the top ten Brazilian Scientific Initiation research selected by the Brazilian Computer Society
4) **SIGIR Student Travel Awards** to present a full research paper at SIGIR'23 in Taipei-Taiwan;
5) **Honorable Mention** in WFA – WebMedia'23;
6) **Honorable Mention** (2nd place) in the Master's Theses Contest (CTDBD) of the **SBBD'21**.

[1]https://scholar.google.com.br/citations?user=TiRmr48AAAAJ&hl=pt-BR

[2]All artifacts can be accessed on https://github.com/waashk/bio-IS

REFERENCES

[1] W. Cunha, F. Viegas, C. França, T. Rosa, L. Rocha, and M. A. Gonçalves, "A comparative survey of instance selection methods applied to non-neural and transformer-based text classification," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–52, 2023.

[2] A. Roy and E. Cambria, "Soft labeling constraint for generalizing from sentiments in single domain," *KBS*, vol. 245, p. 108346, 2022.

[3] A. Ng, "Nuts and bolts of building ai applications using deep learning," *NIPS Keynote Talk*, vol. 64, 2016.

[4] R. Uppaal, J. Hu, and Y. Li, "Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection," *arXiv preprint arXiv:2305.13282*, 2023.

[5] DeepSeek *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[6] W. Cunha, S. Canuto, F. Viegas, T. Salles, C. Gomes, V. Mangaravite, E. Resende, T. Rosa, M. A. Gonçalves, and L. Rocha, "Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling," *Information Processing & Management*, vol. 57, no. 4, p. 102263, 2020.

[7] W. Cunha, V. Mangaravite, C. Gomes, S. Canuto, E. Resende, C. Nascimento, F. Viegas, C. França, W. S. Martins, J. M. Almeida *et al.*, "On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study," *IP&M*, vol. 58, no. 3, p. 102481, 2021.

[8] W. Cunha, A. Pasin, M. Goncalves, and N. Ferro, "A quantum annealing instance selection approach for efficient and effective transformer fine-tuning," in *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, 2024, pp. 205–214.

[9] S. Garcia, J. Derrac, J. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[10] W. Cunha, C. França, G. Fonseca, L. Rocha, and M. A. Gonçalves, "An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification," in *Proceedings of the 46th International ACM SIGIR*, 2023, pp. 665–674.

[11] S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," *PloS one*, 2022.

[12] K. Martins, P. Vaz de Melo, and R. Santos, "Why do document-level polarity classifiers fail?" in *Proceedings of the 2021 Conference of the NAACL: Human Language Technologies*, 01 2021.

[13] W. Cunha, A. Moreo Fernández, A. Esuli, F. Sebastiani, L. Rocha, and M. A. Gonçalves, "A noise-oriented and redundancy-aware instance selection framework," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–33, 2025.

[14] G. Bittencourt, N. Vasconcelos, Y. Andrade, N. Silva, W. Cunha, D. R. Colombo Dias, M. A. Gonçalves, and L. Rocha, "Aware recommender systems (rarss): Recent advances, experimental comparative analysis, discussions, and new directions," *ACM Computing Surveys*, 2025.

[15] C. França, R. C. Lima, C. Andrade, W. Cunha *et al.*, "On representation learning-based methods for effective, efficient, and scalable code retrieval," *Neurocomputing*, 2024.

[16] C. Andrade, F. M. Belém, W. Cunha *et al.*, "On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!"," *IP&M*, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457323000730

[17] B. Zanotto, A. Etges, A. Dal Bosco, E. Cortes, R. Ruschel, S. Martins, A. Souza, C. Valiense, F. Viegas, S. Canuto *et al.*, "Pcv50 automatic classification of electronic health records for a value-based program through machine learning," *Value in Health*, vol. 24, p. S76, 2021.

[18] B. S. Zanotto, A. P. Beck da Silva Etges, A. dal Bosco, E. G. Cortes, R. Ruschel, A. C. De Souza, C. M. V. Andrade, F. Viegas, S. Canuto, W. Luiz, S. Ouriques Martins, R. Vieira, C. Polanczyk, and M. André Gonçalves, "Stroke outcome measurements from electronic medical records: Cross-sectional study on the effectiveness of neural and nonneural classifiers," *JMIR Med Inform*, vol. 9, no. 11, p. e29120, 2021. [Online]. Available: https://medinform.jmir.org/2021/11/e29120

[19] F. Viegas, A. Pereira, W. Cunha, C. França, C. Andrade, E. Tuler, L. Rocha, and M. A. Gonçalves, "Exploiting contextual embeddings in hierarchical topic modeling and investigating the limits of the current evaluation metrics," *Computational Linguistics*, pp. 1–41, 2024.

[20] L. G. Félix, W. Cunha, C. M. de Andrade, M. A. Gonçalves, and J. M. Almeida, "Why are you traveling? inferring trip profiles from online reviews and domain-knowledge," *Online Social Networks and Media*, vol. 45, p. 100296, 2025.

[21] F. Viegas, S. Canuto, W. Cunha, C. França, C. Valiense, G. Fonseca, A. Machado, L. Rocha, and M. Gonçalves, "Pipelining semantic expansion and noise filtering for sentiment analysis of short documents – clusent method," *Journal on Interactive Systems*, vol. 15, no. 1, 2024.

[22] W. Cunha, M. A. Gonçalves, L. Rocha, and G. Dal Bianco, "Mais com menos - processamento de linguagem natural inteligente e sustentável baseado em engenharia de dados e inteligência artificial avançada," *Grandes Desafios da Computaçao no Brasil 2025-2035*, 2024.

[23] G. Fonseca, G. Prenassi, W. Cunha, M. A. Gonçalves, and L. Rocha, "Estratégias de undersampling para redução de viés em classificação de texto baseada em transformers," in *WebMedia*, 2024.

[24] F. Viegas, W. Cunha, C. Gomes, A. Pereira, L. C. da Rocha, and M. A. Gonçalves, "Cluhtm - semantic hierarchical topic modeling based on cluwords," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 8138–8150.

[25] C. França, G. Rabbi, T. Salles, W. Cunha, L. Rocha, and M. A. Gonçalves, "Optimizing tail-head trade-off for extreme multi-label text classification (xmtc) with rag-labels and a dynamic two-stage retrieval and fusion pipeline," in *ACM SIGIR*, 2025.

[26] L. F. Mendes, M. A. Gonçalves, W. Cunha, L. C. da Rocha, T. C. Rosa, and W. Martins, ""keep it simple, lazy" - metalazy: A new metastrategy for lazy text classification," in *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, 2020, pp. 1125–1134.

[27] F. Viegas, S. Canuto, C. Gomes, W. Luiz, T. Rosa, S. Ribas, L. Rocha, and M. A. Gonçalves, "Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling," in *Proceedings of WSDM '19*, 2019, pp. 753–761.

[28] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, W. Cunha, M. A. Gonçalves, and N. Ferro, "Quantumclef 2025-the second edition of the quantum computing lab at clef," in *European Conference on Information Retrieval*. Springer, 2025, pp. 450–458.

[29] C. Andrade, W. Cunha, G. Fonseca, A. Pagano, L. Santos, A. Pagano, L. Rocha, and M. Gonçalves, "Explaining the hardest errors of contextual embedding based classifiers," in *CoNNL'24*, 2024.

[30] J. Costa, G. Oliveira, G. Fonseca, D. Reis, G. Oliveira Teixeira, W. Cunha, L. Rocha, and C. H. G. Ferreira, "Characterizing youtube's role in online gambling promotion: A case study of fortune tiger in brazil," in *Proceedings of the 17th ACM Web Science Conference 2025*, ser. Websci '25, 2025, p. 42–51.

[31] D. Carvalho, A. Pereira, E. Tuler, D. Dias, W. Cunha, and L. Rocha, "Adapting large language models for topic modeling tasks," in *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2024, pp. 954–965. [Online]. Available: https://sol.sbc.org.br/index.php/eniac/article/view/33858

[32] A. P. D. S. Júnior, P. Cecilio, F. Viegas, W. Cunha, E. T. D. Albergaria, and L. C. D. D. Rocha, "Evaluating topic modeling pre-processing pipelines for portuguese texts," in *WebMedia*, 2022, pp. 191–201.

[33] F. Viegas, S. Canuto, W. Cunha, C. França, C. Valiense, L. Rocha, and M. A. Gonçalves, "Clusent–combining semantic expansion and de-noising for dataset-oriented sentiment analysis of short texts," in *WebMedia*, 2023, pp. 110–118.

[34] P. Cecilio, A. Pereira, F. Viegas, J. Rosa, W. Cunha, F. Testa, E. Tuler, and L. Rocha, "Um framework para extração automática de informações em patentes farmacêuticas," in *WebMedia*. SBC, 2023, pp. 97–100.

[35] N. Silva, D. Reis, W. Cunha, E. Tuler, T. Silva, N. Silva, , and L. Rocha, "Integrando avaliações textuais de usuários em recomendação baseada em aprendizado por reforço," in *WebMedia*, 2024.

[36] A. Pasin, W. Cunha, M. A. Gonçalves, N. Ferro *et al.*, "A quantum annealing-based instance selection approach for transformer fine-tuning," in *the 14th Italian Information Retrieval Workshop*, 2022.

[37] L. Félix, W. Cunha, C. Xavier, P. V. de Melo, M. Gonçalves, and J. Almeida, "Geração de roteiros turísticos personalizados em tempo real através de heurística grasp adaptada," in *SBRC*, 2025.

[38] W. Santos, W. Cunha, C. França, G. Fonseca, S. Canuto, L. Rocha, and M. Gonçalves, "Uma metodologia para tratamento do viés da maioria em modelos de stacking via identificação de documentos difíceis," in *SBBD*. SBC, 2023, pp. 408–413.